Relazione Filippo Pitta

assegnazione in data 07/11/2024

Nel presente documento sono descritte le scelte utilizzate al fine di ottenere dei risultati idonei rispetto agli obiettivi preposti.
Gli obiettivi dell'elaborato sono:

1. Pulizia dei dati.
2. Interpolazione spaziale dei dati risultanti.
3. Correlare i risultati con un indice di vigore vegetativo medio, calcolato nell'ultimo anno per quel campo.

Dataset

Il dataset fornito è in formato gpkg formato utilizzato per dati geospaziali. I dati sono suddivisi in 8 features:

- Data/Ora:
- **Longitudin / Latitudine,** delle acquisizioni, CSR è in EPSG:4326
- **RESAKG,** la quantità del prodotto raccolto acquisito, in kg.
- **VELOCITA**, la velocità della mietitrice, in km/h.
- **SUPERFICIE**, la superficie raccolta dall'ultima acquisizione, in mq .
- **UMIDITA'** l'umidità relativa media del prodotto.
- **Geometry:**

Osservazioni sui dati:
data/ora: tempo di acquisizione dei dati 15 secondi, inoltre la stagione dell'acquisizione è avvenuta in estate in data 5 set 2022, questa informazione può essere rilevante.

Analisi dei dati:
È stata fatta una verifica per controllare la consistenza dei dati e il dataset risulta consistente con le feature prese in esame. Sono stati verificati se fossero presenti dei dati mancanti nei campioni, risultano errori di acquisizioni di alcuni dati di "umidità" per 37 campioni. I dati mancanti di "umidità" sono stati colmati con la media dei dati adiacenti.
Sono stati verificati i possibili outlier dai dati quantitativi: VELOCITÀ, RESAKG, UMIDITÀ, AREA.
La visualizzazione dei dati mediante boxplot fig.1 ha fotto emergere molteplici outlier in tutte le features prese in esame. La soluzione adotta per risolvere questo problema è la seguente:

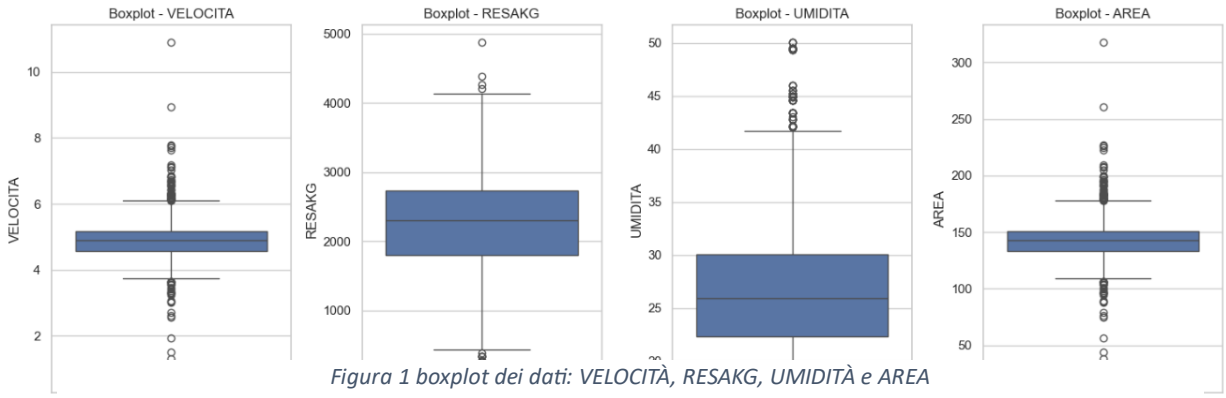


Figura 1 boxplot dei dati: VELOCITÀ, RESAKG, UMIDITÀ e AREA

si è deciso di procedere con il calcolo degli interquartili in quanto si fa utilizzo dei quartili, essi non dipendono dalla media o dalla deviazione standard. Quindi sono stati calcolati gli interquartili per ogni dato quantitativo. Che corrispondono alla differenza del quartile 1 (25%) con il quartile 3 (75%). Successivamente si definiscono i limiti inferiori e superiori. Il numero di dati che compongono il dataset non è estremamente esoso e il numero outlier individuati permettono di applicare un troncamento dei dati.
Il boxplot successivo alla pulizia dei dati fig.2

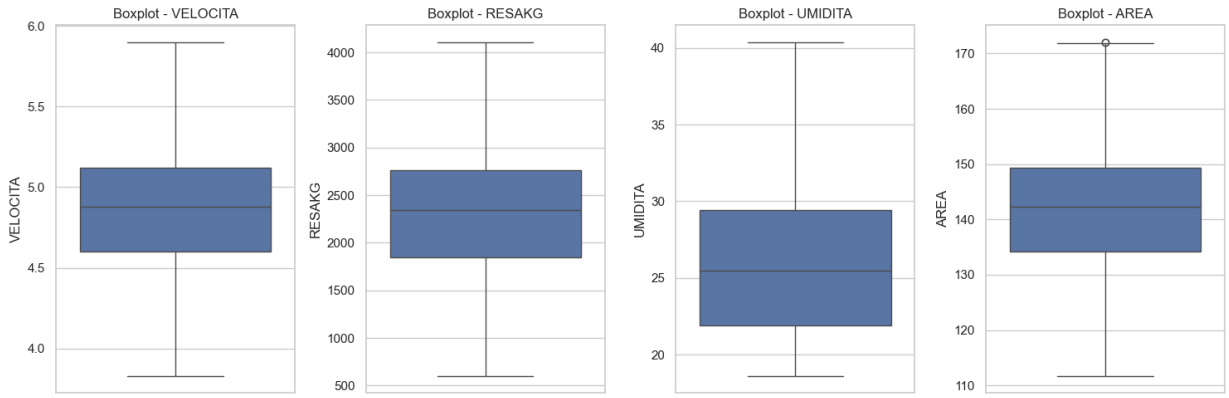


Figura 2 boxplot post pulizia dei dati: VELOCITÀ, RESAKG, UMIDITÀ e AREA

Controllo della distribuzione dei dati

Come controllo preliminare dei dati è preferibile utilizzare un plot per effettuare un analisi visiva dei dati, sono state utilizzate due tipologie differenti di plot, istogramma (fig3) e KDE (Kernel Density Estimation).

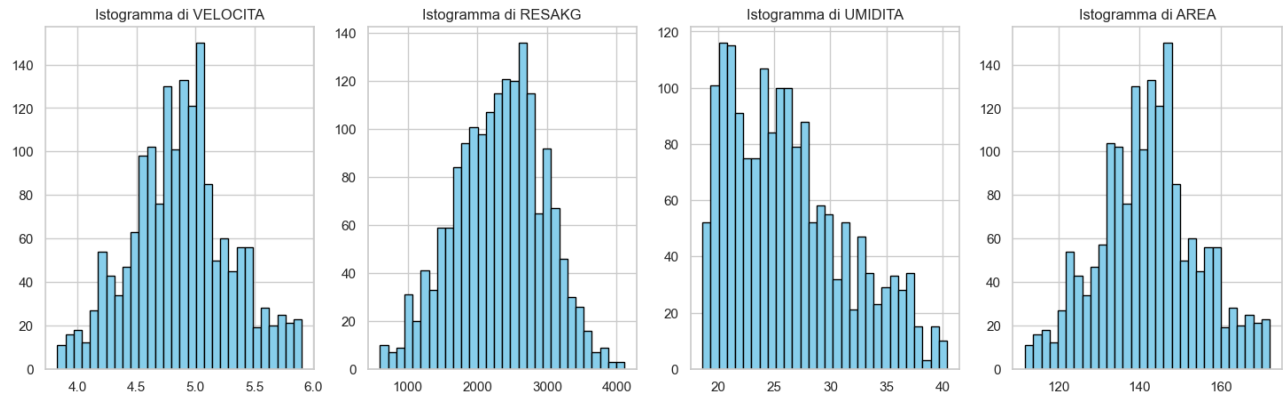
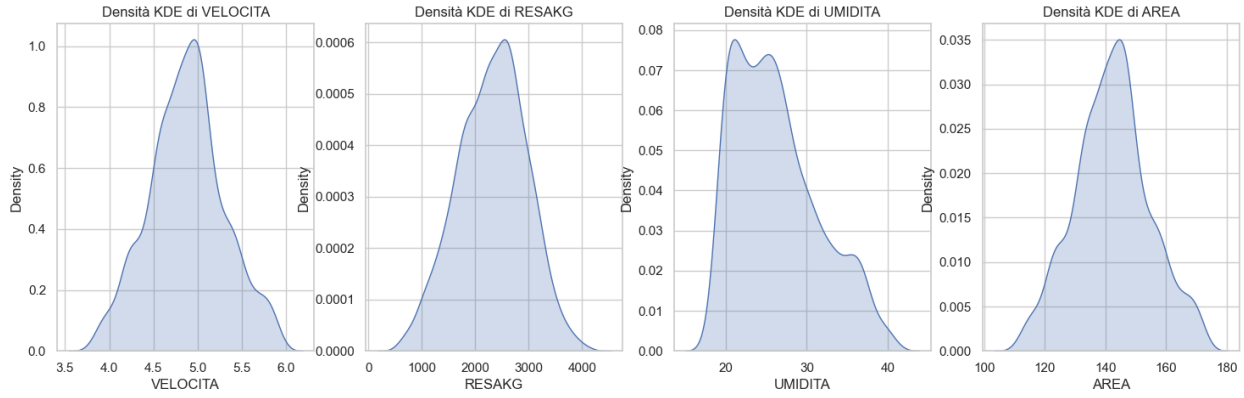


Figura 3 plot degli istogrammi dei dati: VELOCITÀ, RESAKG, UMIDITÀ e AREA



Sono stati analizzati singolarmente i dati per determinare l'andamento delle distribuzioni delle features:

VELOCITA

- La distribuzione è leggermente asimmetrica a destra (positivo), ma il valore è vicino a zero, quindi l'asimmetria è minima.
- La distribuzione ha una forma lievemente più piatta rispetto a una distribuzione normale, ma anche qui il valore è vicino a zero.
- che i dati non seguono una distribuzione normale con un livello di significatività del 5%.

RESAKG

- La distribuzione è leggermente asimmetrica a sinistra (negativo), ma l'asimmetria è bassa.
- Anche in questo caso, la distribuzione è leggermente più piatta della normale.
- Anche per questa variabile, il p-value molto basso suggerisce che la distribuzione non è normale.

UMIDITA

- Questo valore indica una maggiore asimmetria a destra, quindi la distribuzione di UMIDITA è più distorta rispetto alle variabili precedenti.
- La distribuzione è piuttosto piatta rispetto a una distribuzione normale.
- Il basso p-value conferma che anche UMIDITA non segue una distribuzione normale.

AREA

- Come per VELOCITA, c'è una leggera asimmetria positiva, ma è trascurabile.
- Il livello di appiattimento è simile a VELOCITA e non molto diverso dalla distribuzione normale.
- Anche per AREA, il basso p-value suggerisce una distribuzione non normale.

in seguito all'analisi effettuata si è deciso di utilizzare una normalizzazione standard scaler al fine di portare tutti i dati con media 0 e deviazione standard 1.

Dati geospaziali

Nel passo successivo sono stati analizzati i dati geospaziali, si presentano le coordinate di tipo "EPSG:3857" uno standard corrisponde alla proiezione globale, utilizzata per visualizzazione web (fig.4).

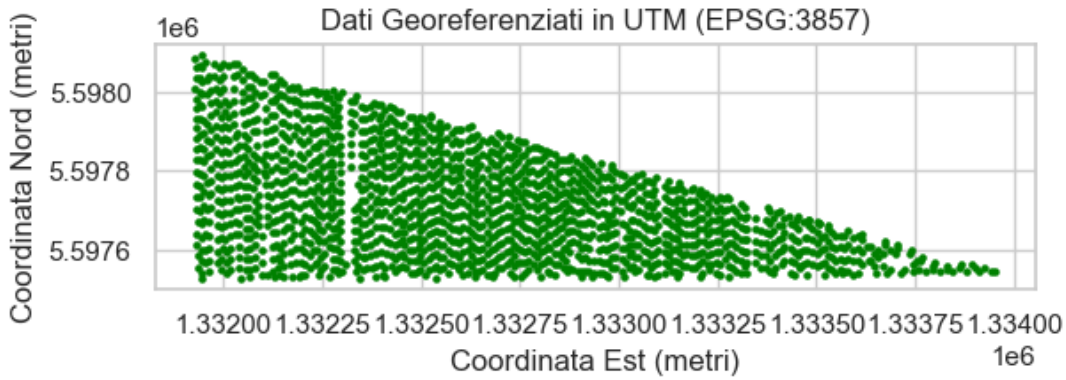


Figura 4 proiezione dati georeferenziali UTM EPSG:3857

Si è passato ad un altro formato “**EPSG:32632”** che è una proiezione locale, specifica per la zona 32N UTM, adatta a mappe più precise in un'area geografica definita fig.5.

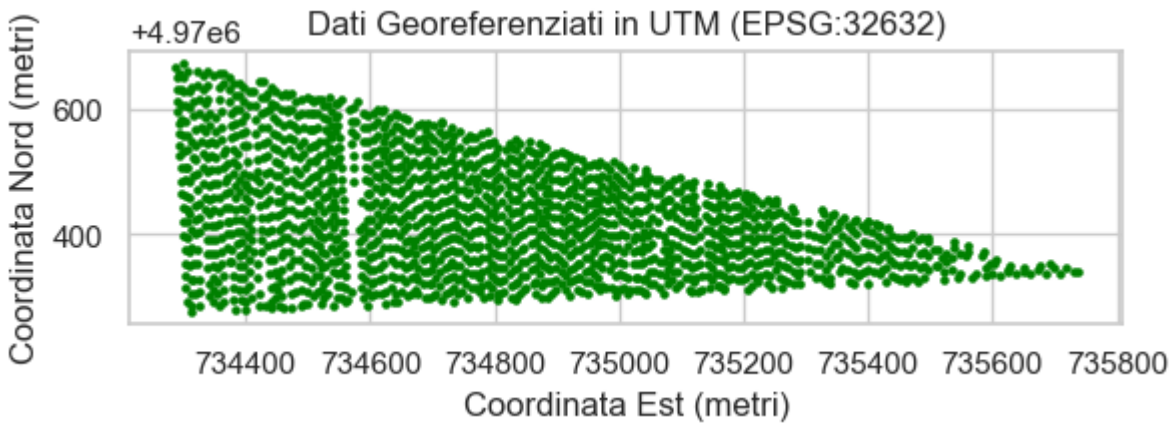


Figura 5 Proiezione dati georeferenziali UTM EPSG:32632

Correlazione tra dati

Sono state osservate le correlazioni tra le diverse variabili interessate. Sono mostrate nella tabella mostrata in fig.6

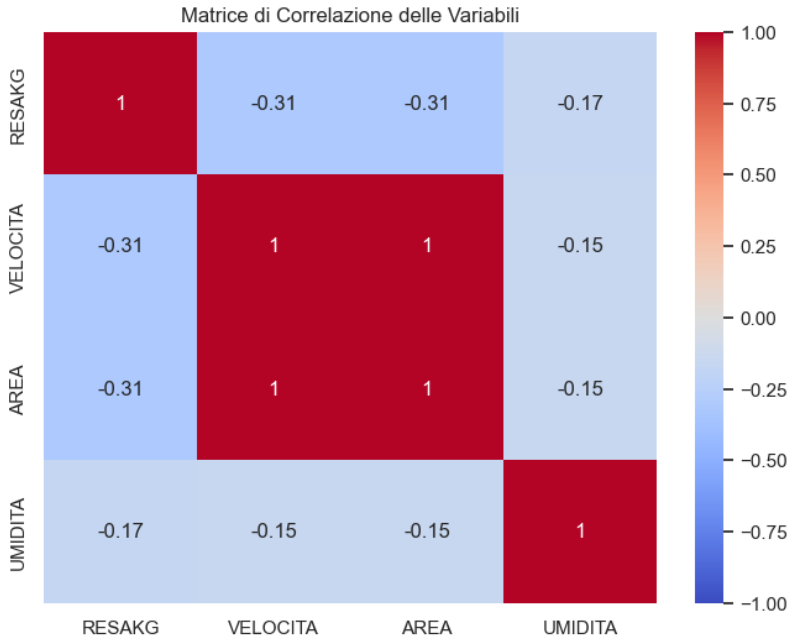


Figura 6 Matrice di Correlazione tra Variabili

Correlazione tra RESAKG e altre variabili:

- La correlazione tra RESAKG e VELOCITA è moderatamente negativa (-0,31), indicando che una velocità maggiore potrebbe essere associata a una minore resa. Questo potrebbe suggerire che a velocità più elevate la raccolta potrebbe essere meno efficace.
- La correlazione tra RESAKG e AREA è anch’essa -0,31, suggerendo una relazione simile, sebbene non fortemente significativa. Forse aree più grandi raccolte in un singolo passaggio potrebbero influire sull’efficienza.

- La correlazione tra RESAKG e UMIDITA è solo leggermente negativa (-0,17), indicando una correlazione molto debole e probabilmente non significativa.

Correlazione tra VELOCITA e AREA:

- C'è una correlazione perfetta (1) tra VELOCITA e AREA, il che è insolito e potrebbe derivare da qualche aspetto del dataset o della raccolta dati. Potrebbe indicare che la velocità e l'area misurate sono direttamente proporzionali.

Correlazione tra altre variabili e UMIDITA:

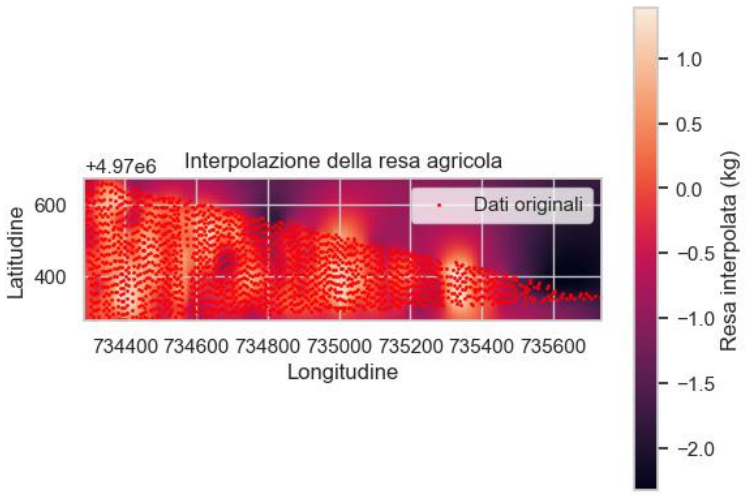
- UMIDITA ha una correlazione molto bassa con le altre variabili, suggerendo che la variazione dell'umidità non è legata a VELOCITA, AREA o RESAKG in modo significativo.

Interpolazione spaziale dei dati

È stata eseguita l'interpolazione spaziale sui dati tramite il metodo di Kriging. Questo approccio è stato scelto in quanto il Kriging permette di ottenere stime ottimali dei valori spaziali, tenendo conto delle correlazioni spaziali tra i dati.

Per ogni variabile di interesse: VELOCITÀ, RESAKG, UMIDITÀ, AREA, sono state estratte le coordinate geografiche dei punti di misura e i relativi valori. Successivamente, è stata definita una griglia regolare di interpolazione, sulla quale sono stati calcolati i valori stimati per ogni punto usando il Kriging ordinario. L'output dell'interpolazione è stato salvato come immagine raster in formato GeoTIFF, che rappresenta la distribuzione spaziale della variabile stimata.

In questo processo, è stato scelto un modello di variogramma esponenziale per descrivere la struttura spaziale dei dati. Tuttavia, è possibile utilizzare anche altri modelli in base alla natura dei dati stessi. Questo tipo di interpolazione fornisce una mappa continua dei valori di una variabile, che è particolarmente utile per visualizzare e analizzare fenomeni spaziali distribuiti in modo non uniforme.



Calcolo della resa NDVI

Si è pensato di creato sinteticamente dei dati approssimativi ma con esito negativo rispetto all'obiettivo. Un'idea per riuscire a prendere i dati richiesti mediante i satelliti di google earth in particolare il satellite "Copernicus" che permette di estrarre dati per poter calcolare gli indici NDVI (https://developers.google.com/earth-engine/datasets/catalog/sentinel-2).